

Vocabulary Richness

Stefan Evert, Fotis Jannidis, Thomas Proisl

Kallimachos Treffen, 19.1.2018

Aspekte der Komplexität literarischer Texte

- Stil
 - Vokabular
 - Syntax
 - Register
 - Uneigentliche Rede
- Ästhetik
 - Beziehung von Inhalt und Form
- Textwelt
 - Komplexität der Figurenkonstellation
 - Art und Vielfalt der dargestellten Emotionen
- Symbolische Elemente der Textwelt
- Fülle der Intertextualität
- Polyvalenz des Textes, Fülle von möglichen Interpretationen

Vocabulary richness, lexical diversity

- Nicht ‚readability‘: („ease of reading, especially as the result of the writing style“). Eigenes Forschungsfeld
- ‚lexical diversity‘ oder ‚diversity of vocabulary‘ (Carroll 1938)
‚lexical richness‘ oder ‚vocabulary richness‘ (Yule 1944)
- Grundidee: die gemessene LD ist ein Indikator für die Größe des Vokabulars einer Person

Grundlegende Idee

- „Und dann weinte er und weinte und weinte. Und dann war er ruhig.“

- und 4

- weinte 3

- er 2

type-token ratio (Johnson 1939): 5 /13

- dann 2

- war 1

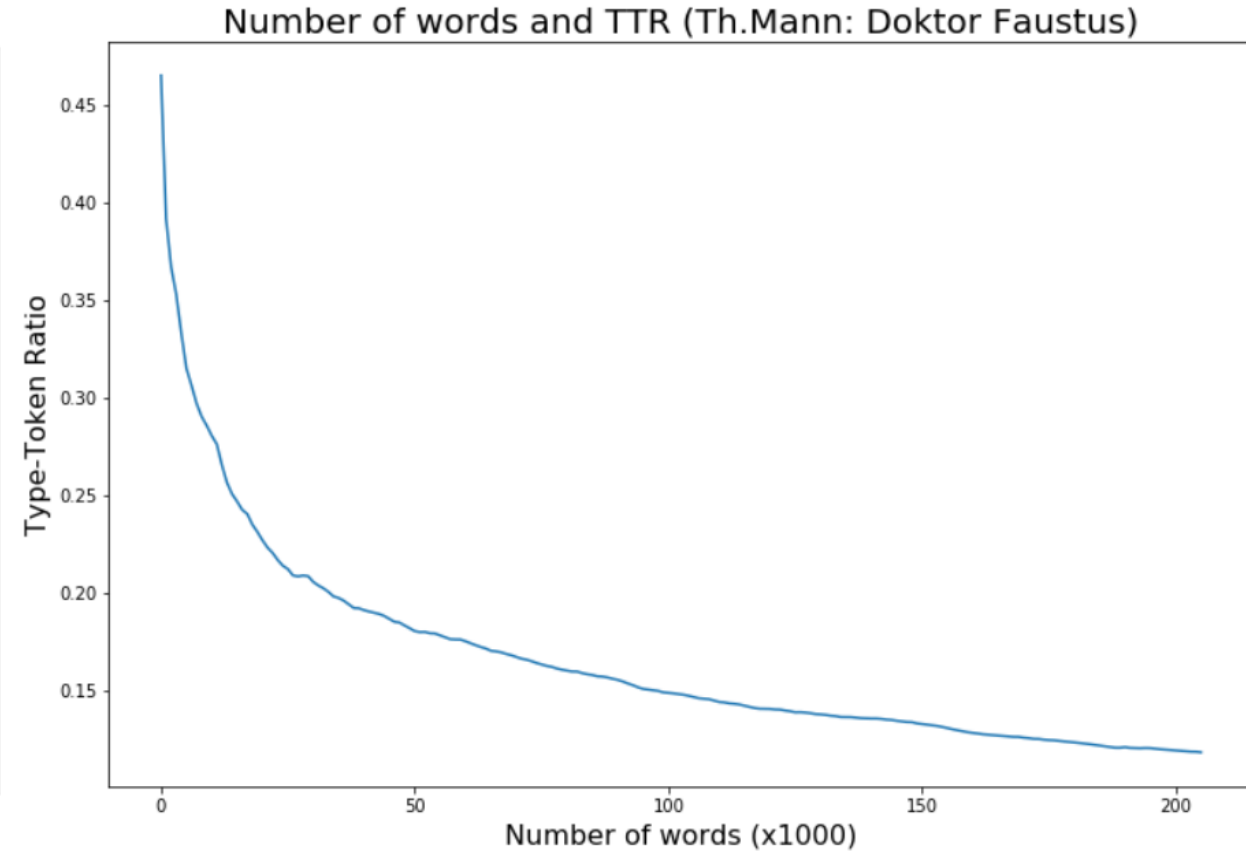
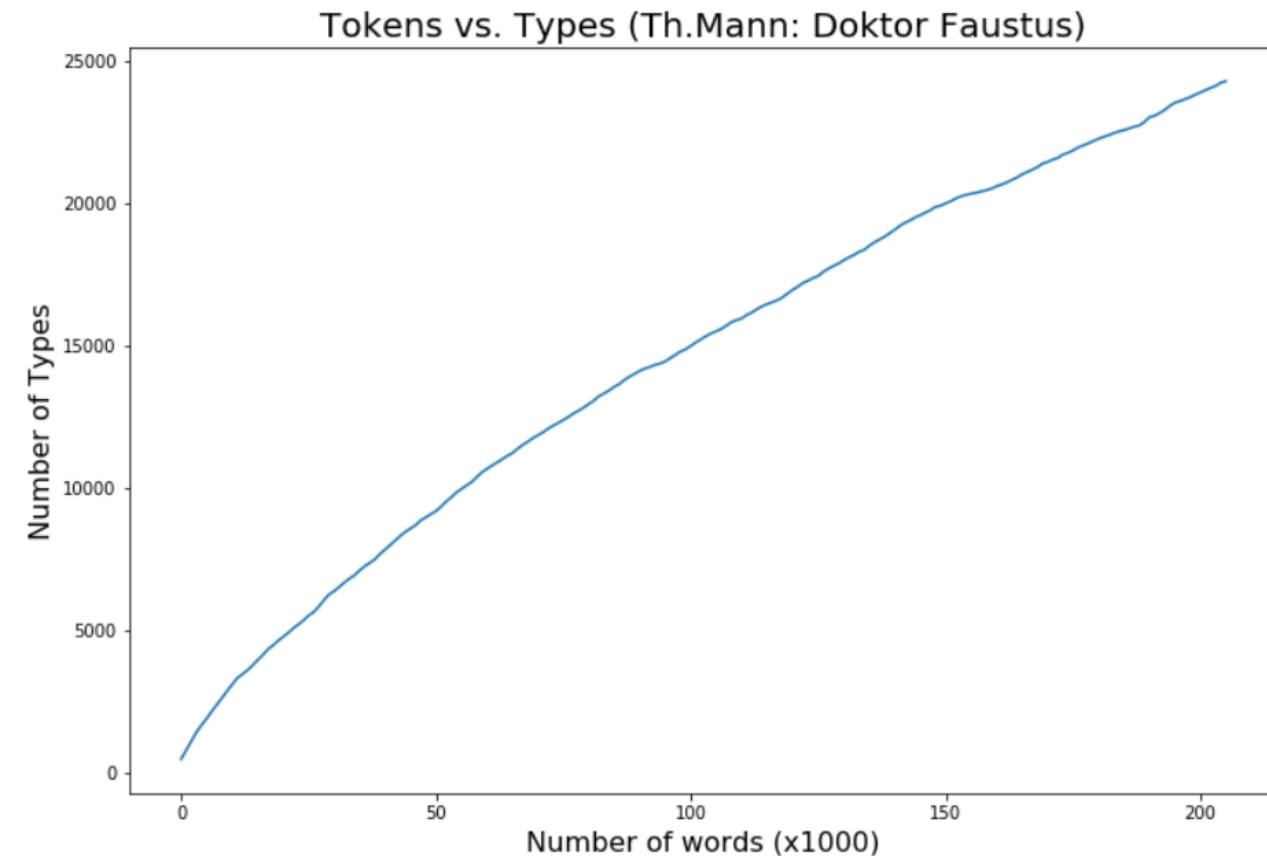
- ruhig 1

V = 5 ‚Types‘

N = 13 ‚Tokens‘

Das Problem

- Fast alle LD Maße variieren abhängig von der Größe des Textausschnitts



LD als Maß einer Texteigenschaft

- LD Maße sollen ein Maß für eine charakteristische Texteigenschaft sein (Tanaka-Ishii, Aihara 2015)

Aber:

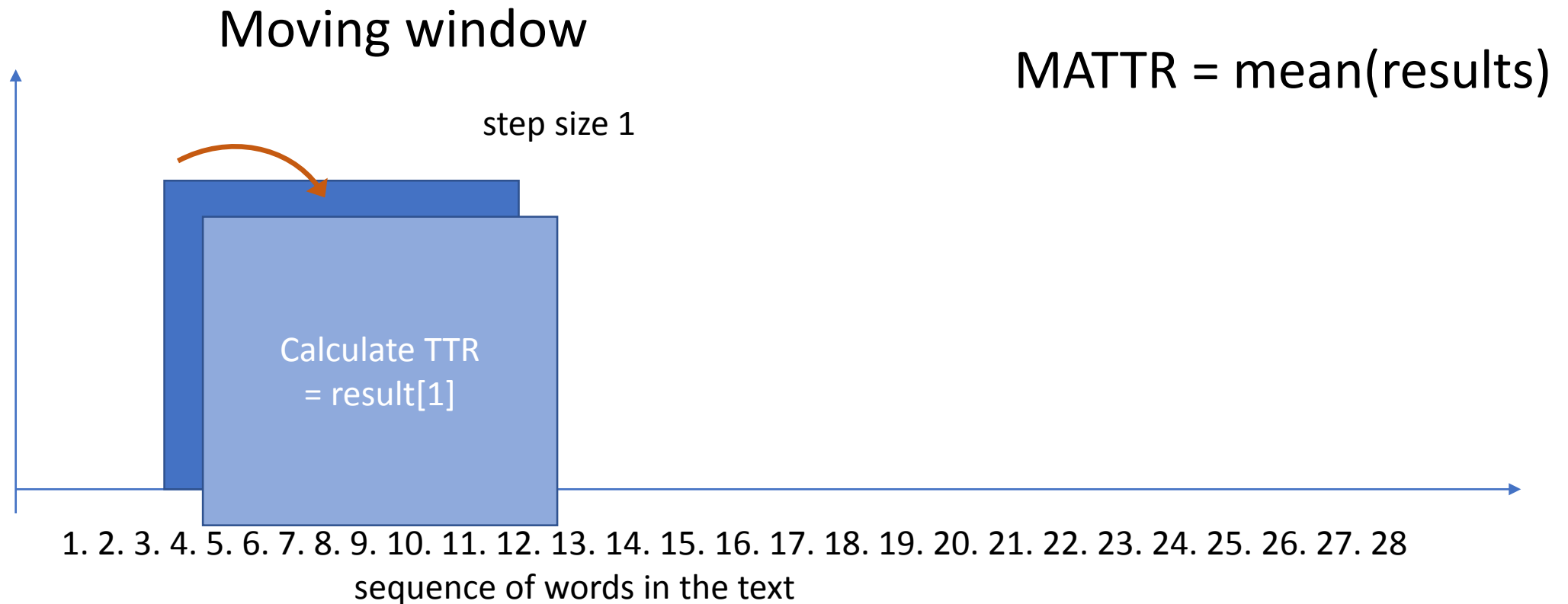
- Fast alle vorgeschlagenen Maße hängen in hohem Grad von der Textlänge ab (Tweedie, Baayen 1998)

Fülle verschiedener Maße

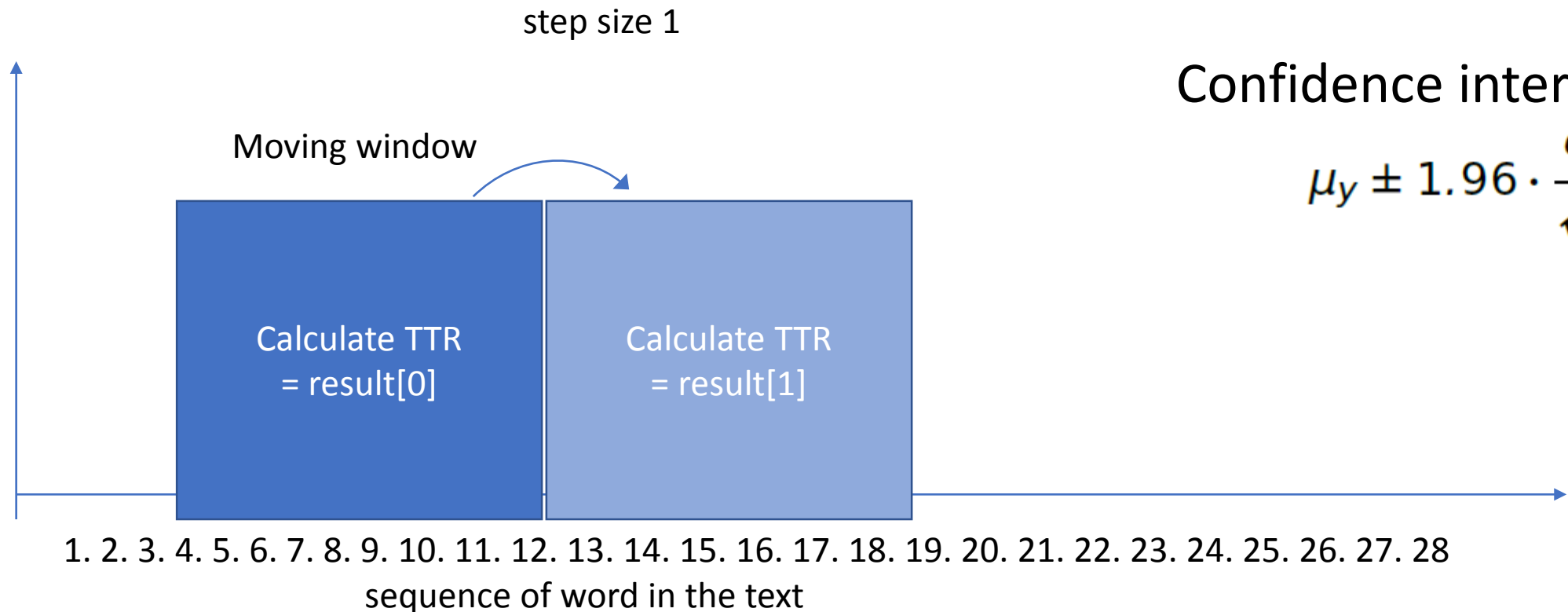
- Type-token ratio
- Brunet's W ,
- Carroll's CTTR,
- Dugast's k ,
- Entropy,
- Guiraud's R ,
- Sichel's S ,
- Tuldava's LN ,
- Honoré's H ,
- Herdan's C ,
- HD-D,
- Michéa's M ,
- Summer's S ,
- Dugast's U ,
- Herdan's V_m ,
- MTLD,
- Maas' a^2 ,
- Simpson's D ,
- Yule's K

The Moving-Average Type-Token Ratio (MATTR)

Covington & McFall (2010)



Standardized Type-Token Ratio (STTR)



STTR = mean(results)

Confidence interval

$$\mu_y \pm 1.96 \cdot \frac{\sigma_y}{\sqrt{n}}$$

MTLD

McCarthy, Jarvis 2010

- **„of the people, by the people, for the people, shall not perish from the earth“**
- Für jedes Wort-Token wird inkrementell ein type-token Wert (TTR) berechnet: „... *of* (1.00) *the* (1.00) *people* (1.00) *by* (1.00) *the* (.800) *people* (.667) *for* (.714) *the* (.625) *people* (.556) *shall* (.600) ...“
- Wenn ein Schwellenwert erreicht wird (zumeist .720), wird der Faktorzähler erhöht und der TTR-Wert zurückgesetzt:
„... *of* (1.00) *the* (1.00) *people* (1.00) *by* (1.00) *the* (.800) *people* (.667) *for* (.714) *the* (1.00) *people* (1.00) *shall* (1.00) ...“

MTLD

- $fc_1 = \text{factor_count}(w_0, w_1, \dots, w_n)$
- $fc_2 = \text{factor_count}(w_n, w_{n-1}, \dots, w_1)$
- $MTLD = \text{mean}(fc_1 + fc_2)$

MTLD kann also verstanden werden als die durchschnittliche Anzahl an Worten, bis der TTR-Schwellenwert (.720) erreicht wird.

Methodologische Aspekte

Kernfragen für die erste Projektphase:

- Wie zuverlässig sind diese Maße? (→ Konfidenzintervall)
Wann ist ein Unterschied zwischen zwei Texten, ... signifikant?
- Was messen wir eigentlich?
Welche der zahlreichen quantitativen Maße sind sinnvoll?
- SE: Wie hängen Produktivität und LD zusammen?

Zweite Projektphase: Erweiterung auf semantische Komplexität

Signifikanz von LD

- Viele Arbeiten nur deskriptiv, keine statistische Auswertung
- State of the art: **LNRE-Modelle** auf Basis des Zipfschen Gesetzes (Baayen 2001, Evert 2004, Evert & Baroni 2007)
 - parametrisches Modell ermöglicht Inferenz: Stichprobe → Grundgesamtheit

Aber zahlreiche Probleme ...

- Zipfsches Gesetz zu simpel (Baayen 2001, Kornai 1999)
- Nichtzufälligkeit von Texten (Baroni & Evert 2007)
- Parameterschätzung der LNRE-Modelle nicht robust (Evert 2017)

Signifikanz: Bootstrapping / Cross-validation

- Nichtparametrische Inferenz: Bootstrapping
- Nicht auf LD anwendbar, da Type-Token-Statistiken verfälscht werden

Unser Ansatz (Evert/Wankerl/Nöth 2017):

- Kombination von **Cross-validation** und Bootstrapping
- Jeder Text wird in gleich große Schnipsel (*fold*s) zerlegt, und das gewünschte LD-Maß auf jedem *fold* bestimmt: y_1, \dots, y_k
- Globalwert für gesamten Text:

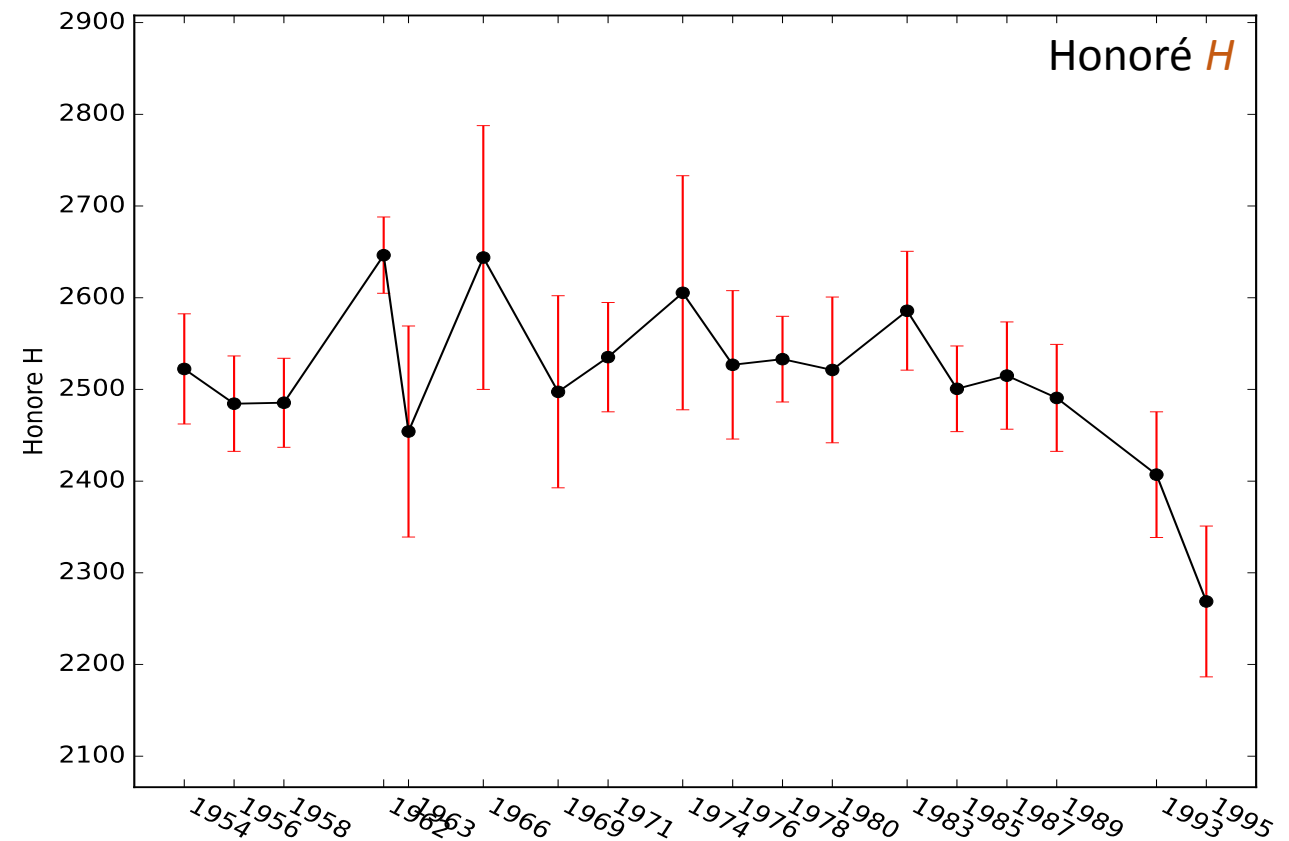
$$\sigma_{\bar{y}} = \frac{\sigma_y}{\sqrt{k}}$$

$$\bar{y} = \frac{y_1 + \dots + y_k}{k}$$

analog zu STTR, aber
auch auf alle anderen
LD-Maße anwendbar
(inkl. MTLD)

Eine erste Fallstudie ...

- Sinkende LD in Romanen von Iris Murdoch als Frühsymptom ihrer Alzheimer-Erkrankung? (Evert/Wankerl/Nöth 2017)
- Frühere Ergebnisse widersprüchlich (Garrard et al. 2005; Le et al. 2011)
- Unterschiede zwischen Texten wurden nicht auf statistische Signifikanz getestet
- CV/Bootstrapping erweist sich als vielversprechend



Forschungsfragen

- Welche Texte unterscheiden sich signifikant bzgl. LD?
 - Welche Textgröße ist für zuverlässige Aussagen erforderlich?
 - Wie hängen LD-Maße (inkl. STTR) von *fold*-Größe ab?
 - Sind Ergebnisse für unterschiedliche *fold*-Größen konsistent?
 - Sind Ergebnisse verschiedener LD-Maße konsistent?
- Was messen wir eigentlich?
- Gibt es die „wahre“ LD?
 - Oder erfassen verschiedene Maße unterschiedliche Aspekte von LD?

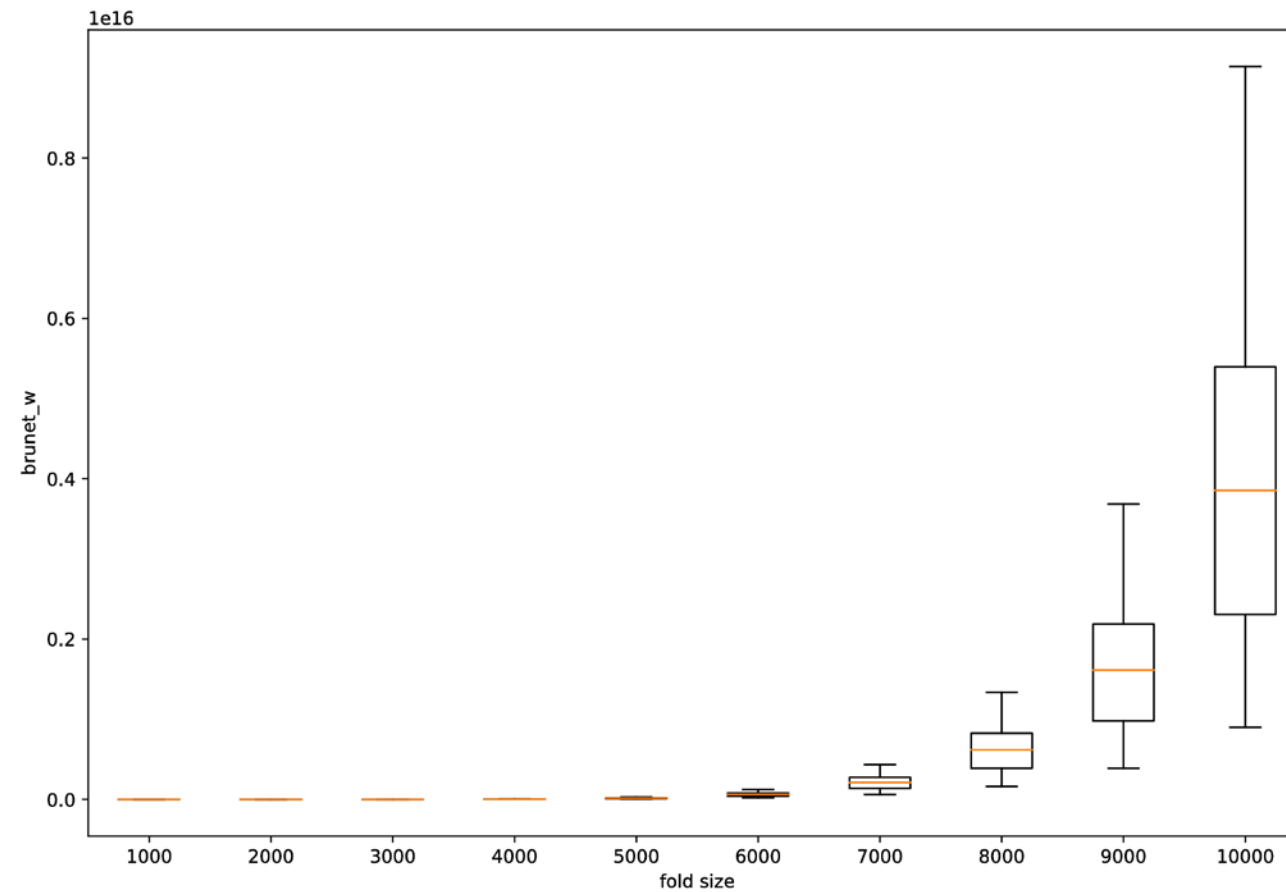
Beschreibung (1)

- Fragestellungen
 - Ist das Verhalten der Maße abhängig von der Textlänge?
 - Es ist unklar, was die Maße eigentlich messen. Können wir zumindest feststellen, welche Maße etwas ähnliches messen?
- Datenbasis
 - Delta-Korpora in drei Sprachen (deutsch, englisch, französisch; je 25 3 Romane)
 - Gutenberg-Korpus (deutsche Romane aus dem langen 19. Jahrhundert)

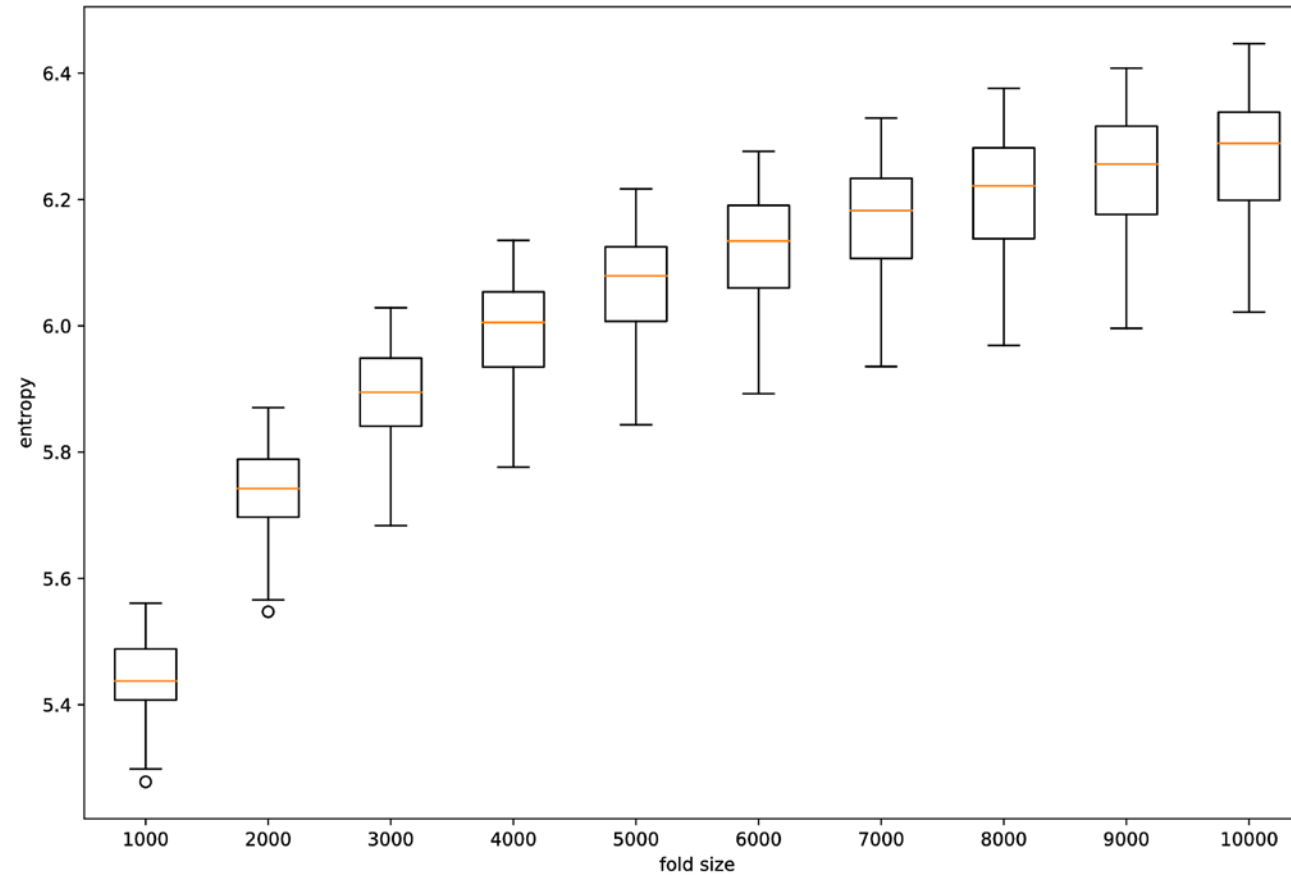
Beschreibung (2)

- Methode
 - Zerlegung der Texte in Abschnitte (Folds) gleicher Länge (Fold-Size)
 - Fold-Size: 1.000–10.000 Wörter
 - Ignorieren überzähliger Wörter am Textende
 - Berechnung von Komplexitätsmaßen für jeden Abschnitt
 - Mittelwerte und Konfidenzintervalle für jeden Text
- Konkretisierung der Fragestellungen
 - Sind die Maße abhängig von der Fold-Size?
 - Längere Abschnitte ! größere/kleinere Werte?
 - Ist das Ranking der Texte abhängig von der Fold-Size?
 - $A > B$ für kleine Fold-Size, aber $B > A$ für große Fold-Size?
 - Rangkorrelationen zwischen Fold-Sizes
 - Gibt es Gruppen von Maßen, die dasselbe messen?
 - Sehr starke bis perfekte (Rang-)Korrelation zwischen Maßen?

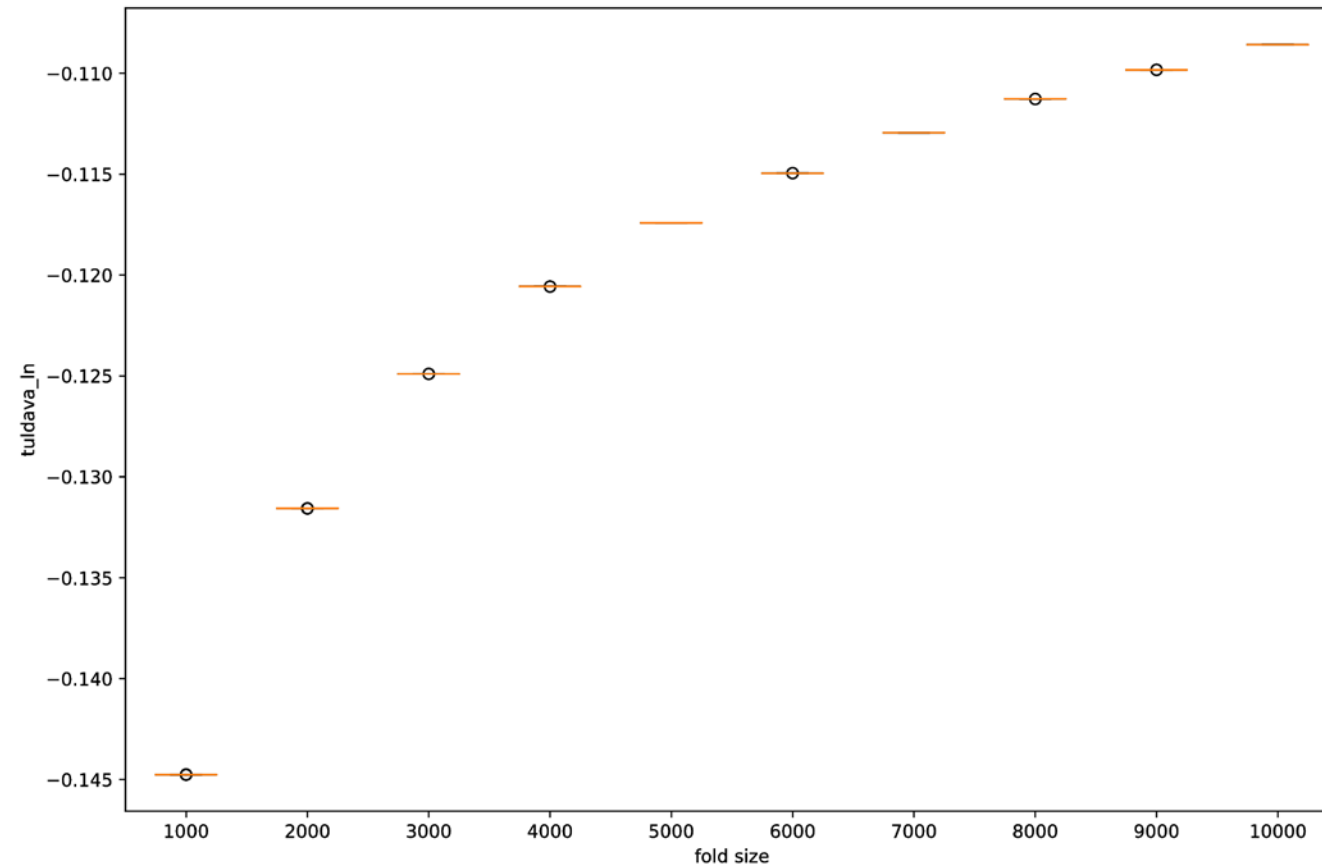
Ergebnisse – Stabilität der Werte (Brunet's W)



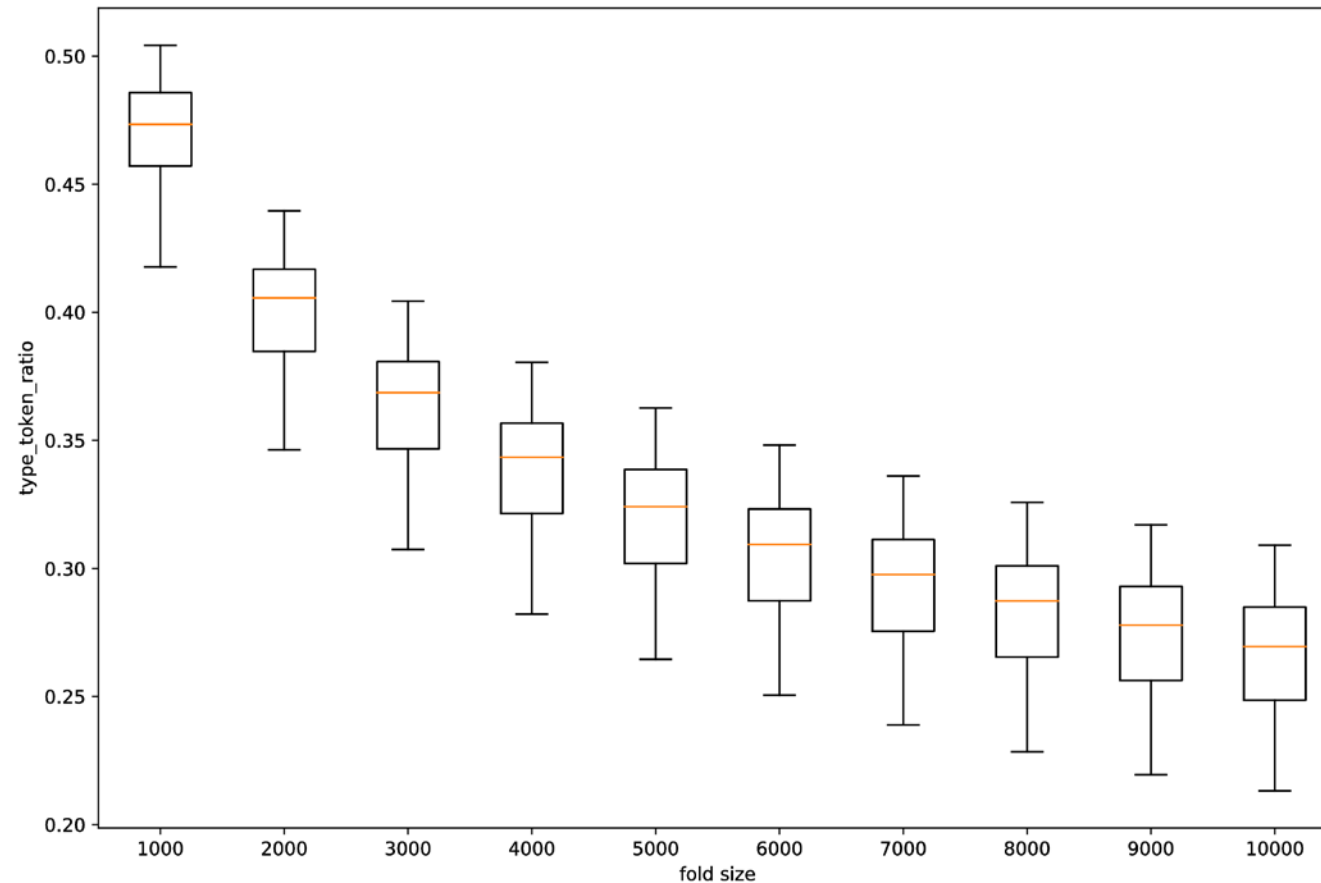
Ergebnisse – Stabilität der Werte (Entropy)



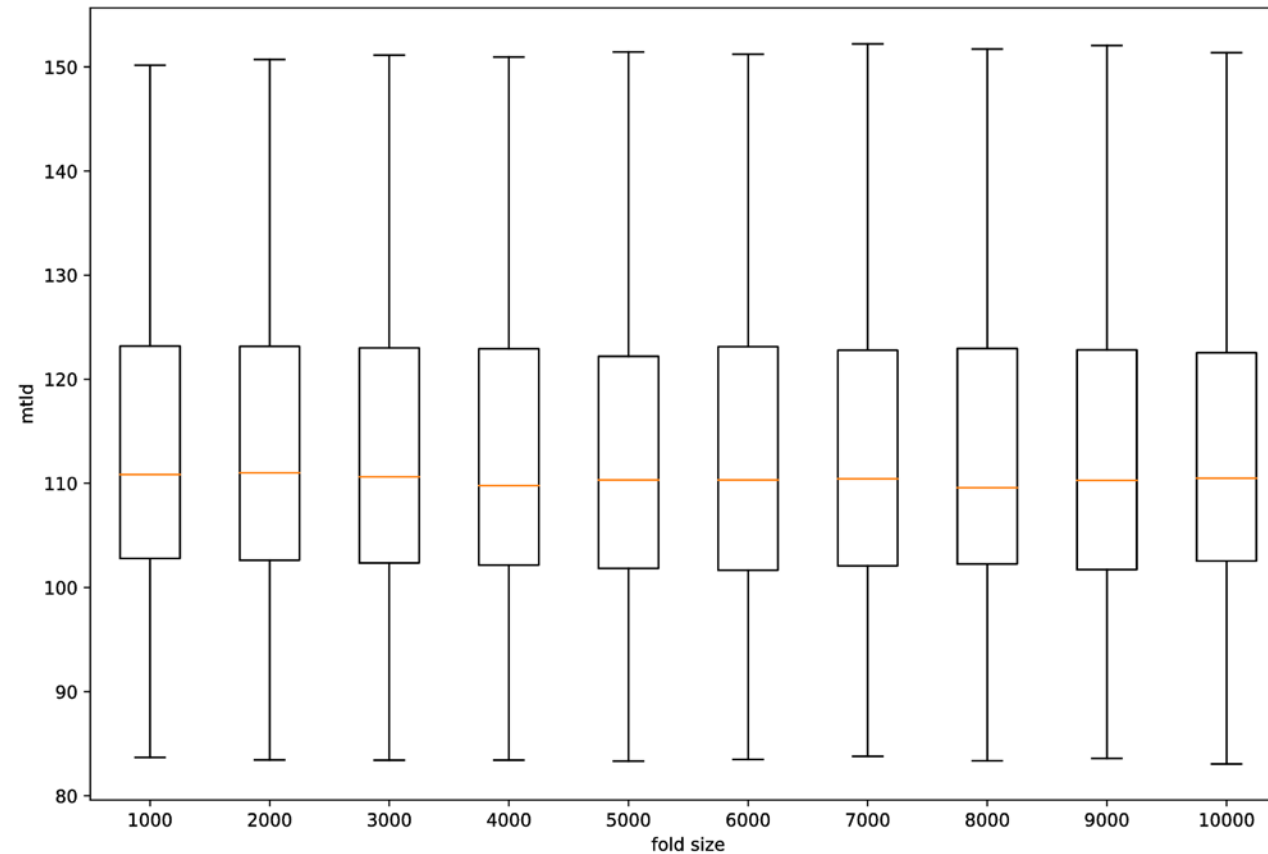
Ergebnisse – Stabilität der Werte (Tuldava's LN)



Ergebnisse – Stabilität der Werte (Type-token ratio)



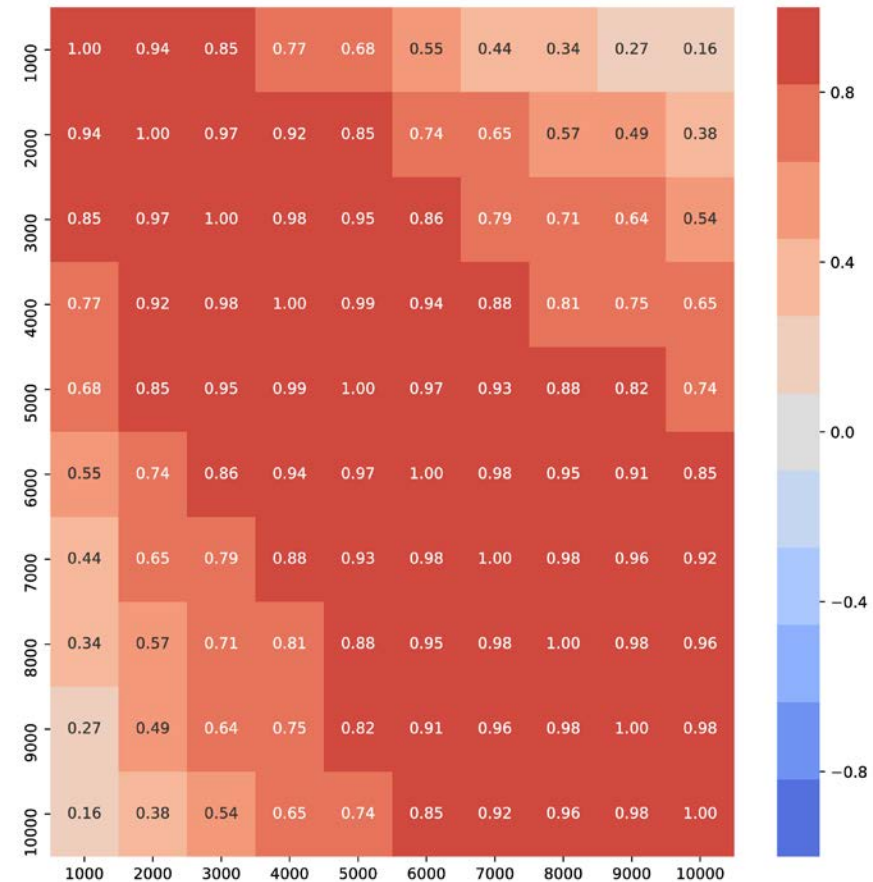
Ergebnisse – Stabilität der Werte (MTLD)



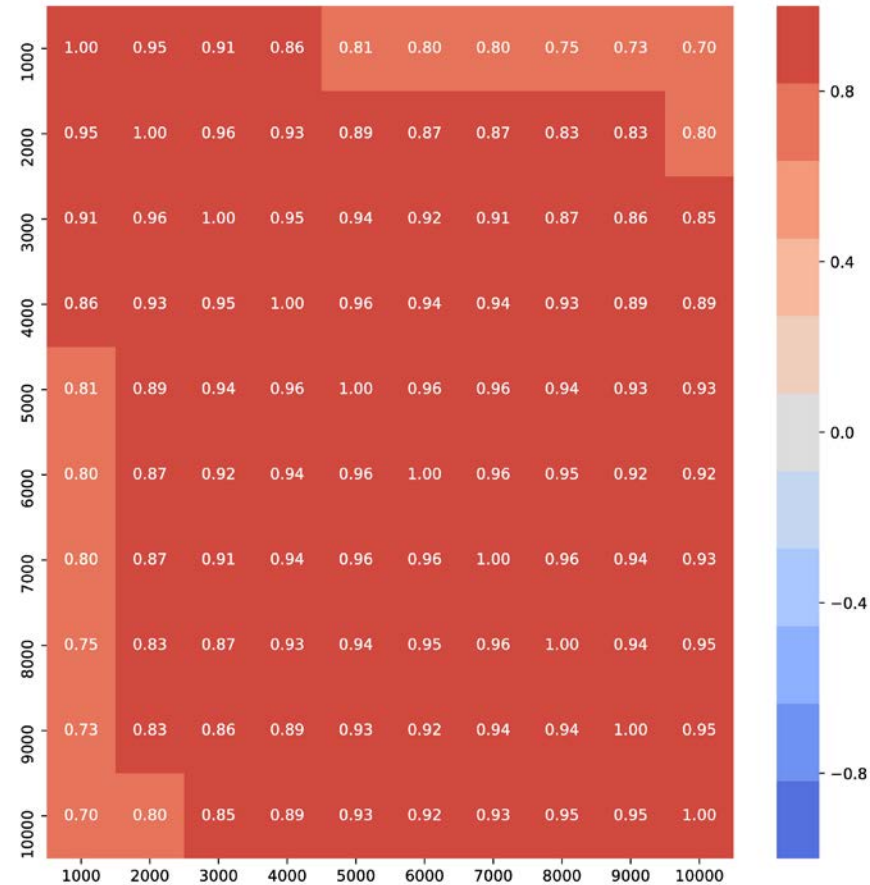
Ergebnisse – Stabilität der Werte

- Werden größer: Brunet's W, Carroll's CTTR, Dugast's k, Entropy, Guiraud's R, Sichel's S, Tuldava's LN, (Honoré's H)
- Werden kleiner: Herdan's C, HD-D, Michéa's M, Summer's S, Type-token ratio
- Stabil: Dugast's U, Herdan's Vm, MTLD, Maas' a_2 , Simpson's D, Yule's K

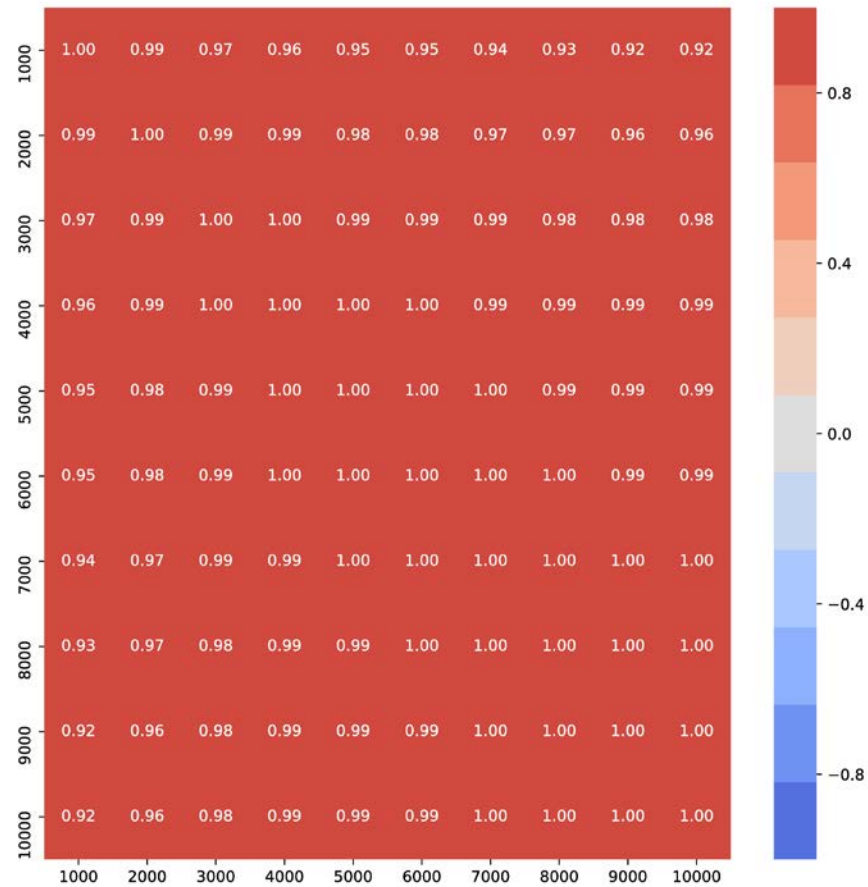
Ergebnisse – Stabilität der Rankings (HD-D)



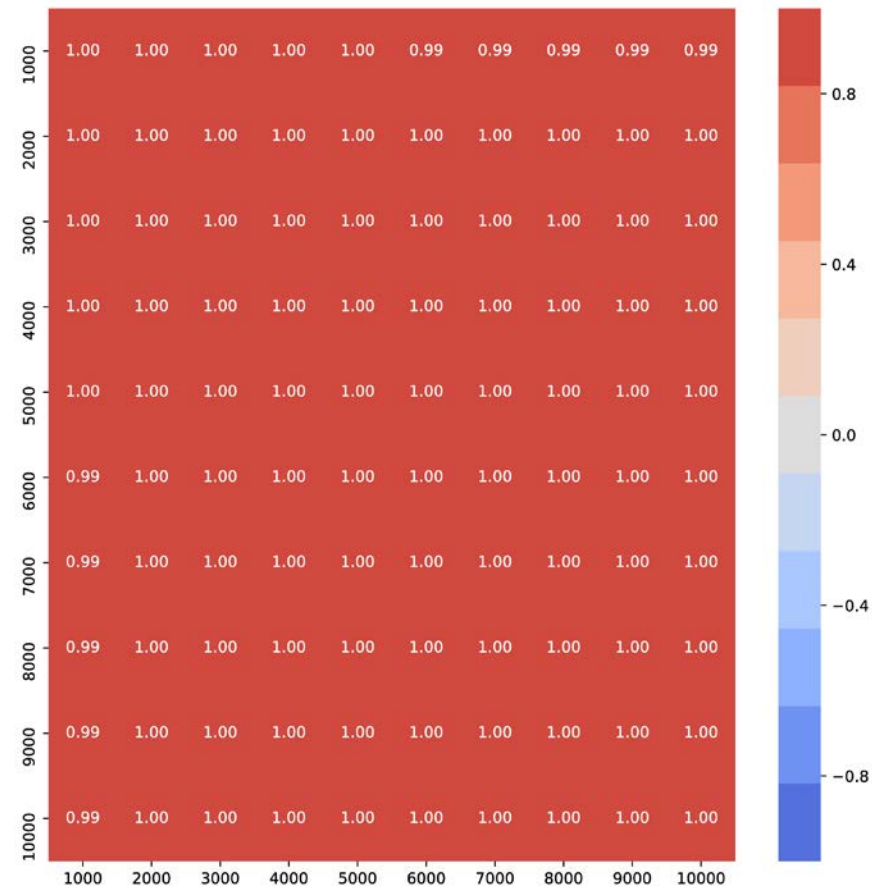
Ergebnisse – Stabilität der Rankings (Sichel's S)



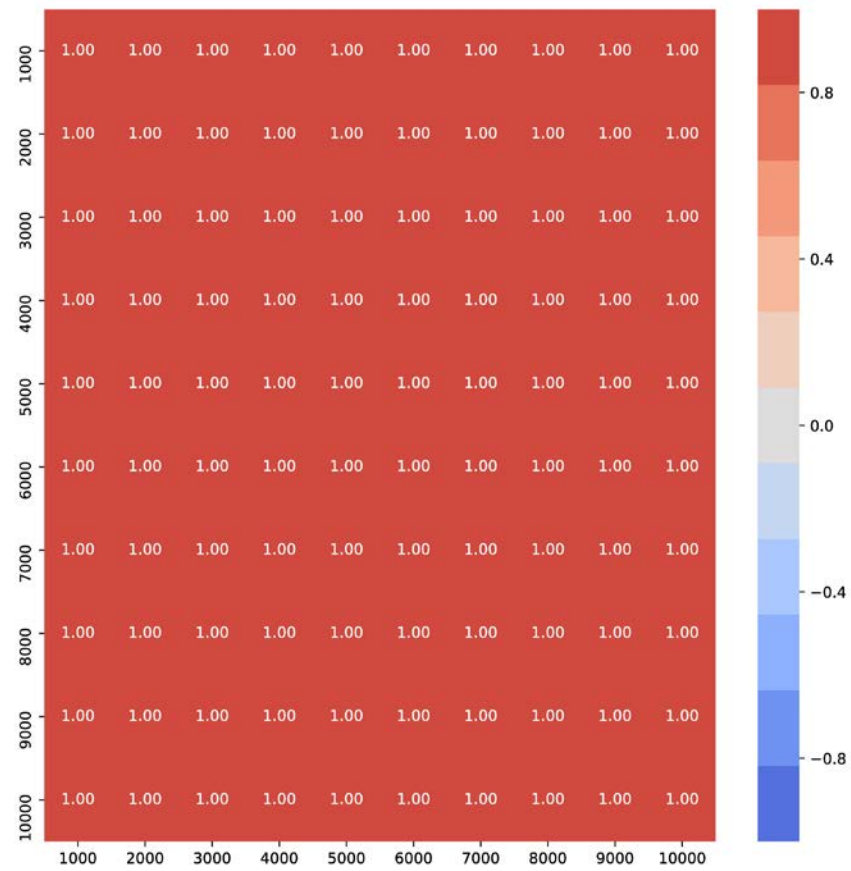
Ergebnisse – Stabilität der Rankings (Brunet's W)



Ergebnisse – Stabilität der Rankings (Yule's K)

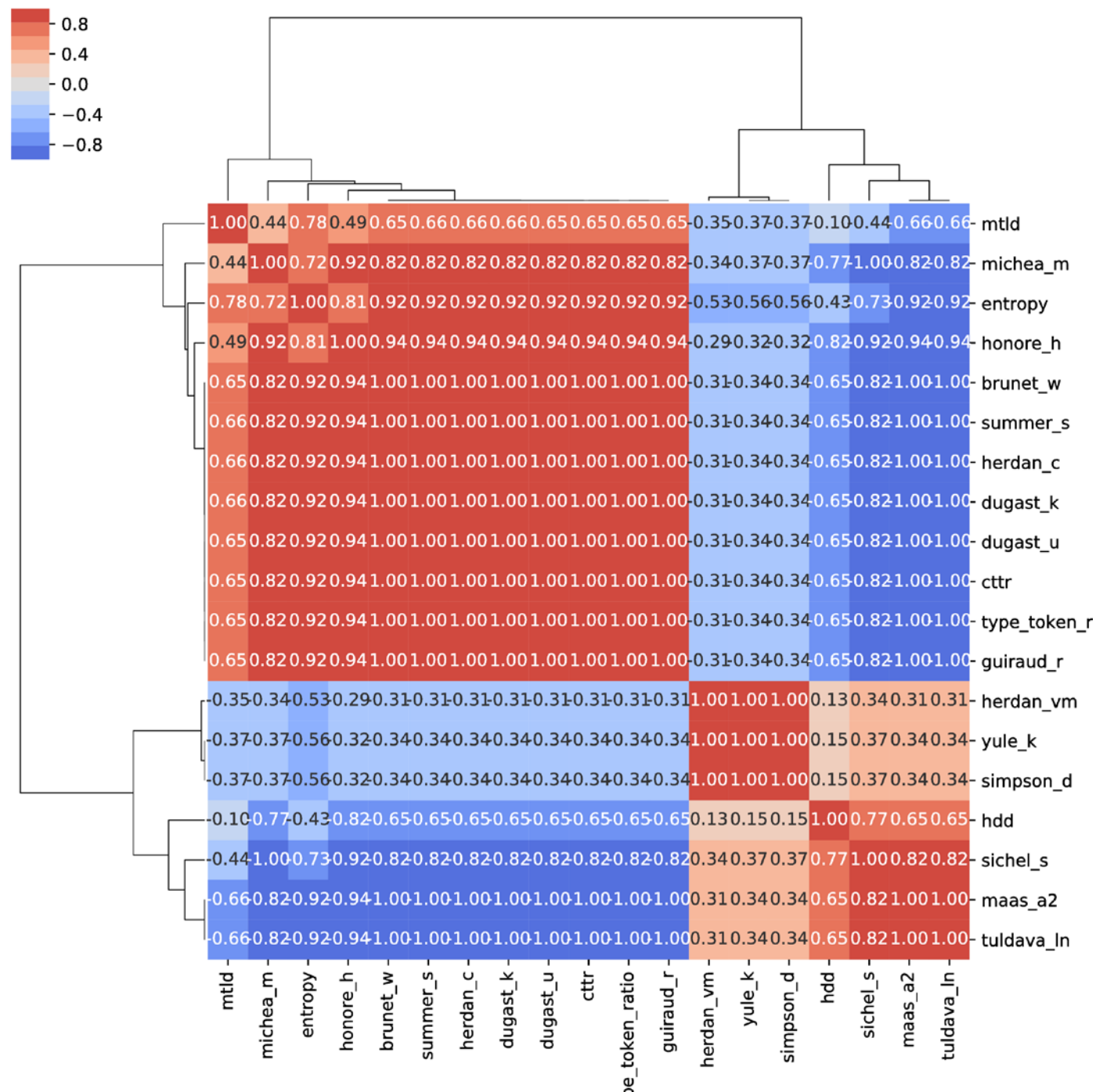


Ergebnisse – Stabilität der Rankings (Herdan's V_m)



Ergebnisse – Stabilität der Rankings

- Nicht stabil: HD-D, Michéa's M, Sichel's S
- Einigermaßen stabil: Brunet's W, Carroll's CTTR, Dugast's k, Dugast's U, Entropy, Guiraud's R, Herdan's C, Maas' a_2 , Summer's S, Tuldava's LN, Type-token ratio, (Honoré's H)
- Stabil: Herdan's V_m , MTLD, Simpson's D, Yule's K



Ergebnisse – Korrelationen zwischen den Maßen

- Perfekt: Herdan's V_m , Simpson's D , Yule's K
- Perfekt: Maas' a^2 , Tuldava's LN
- Perfekt: Brunet's W , Carroll's $CTTR$, Dugast's k , Dugast's U , Guiraud's R , Herdan's C , Summer's S und Type-token ratio
- Perfekt negativ: Brunet's W etc. und Maas' a^2 etc.
- Perfekt negativ: Michéa's M und Sichel's
- Stark: Entropy und Honoré's H mit Brunet's W etc., aber nicht so stark miteinander
- Stark: Michéa's M und Sichel's mit Honoré's H , nicht so stark mit Brunet's W etc. und Maas' a^2 etc.
- MTLD bildet eigene Gruppe, am ähnlichsten zu Entropy
- HD-D bildet eigene Gruppe

Fazit

- Fünf Gruppen von Maßen
 - HD-D: Werte und Ranking instabil
 - MTLD: Werte und Ranking stabil
 - Michéa's M und Sichel's S: Werte und Ranking instabil
 - Herdan's Vm, Simpson's D, Yule's K: Werte und Ranking stabil
 - Brunet's W, Carroll's CTTR, Dugast's k, Dugast's U, Guiraud's R, Herdan's C, Summer's S und Type-token ratio; Maas' α^2 , Tuldava's LN; (Entropy und Honoré's H): Werte teilweise stabil (Dugast's U, Maas' α^2), Ranking einigermaßen stabil

Literatur

- Baayen, R. Harald (2001). *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht.
- Baroni, Marco & Evert, Stefan (2007). Words and echoes: Assessing and mitigating the non-randomness problem in word frequency distribution modeling. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 904–911, Prague, Czech Republic.
- Covington, M.A., J.D. McFall (2010). Cutting the Gordon Knot. *Journal of Quantitative Linguistics* 17,2, 94-100.
- Evert, Stefan (2004). A simple LNRE model for random character sequences. In *Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2004)*, pages 411–422, Louvain-la-Neuve, Belgium.
- Evert, Stefan (2017a). Measures of productivity and lexical diversity. *Poster at the ICAME 38 Conference*, Prague, Czech Republic.
- Evert, Stefan & Baroni, Marco (2007). zipfR: Word frequency distributions in R. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions*, pages 29–32, Prague, Czech Republic.
- Evert, Stefan; Wankerl, Sebastian; Nöth, Elmar (2017). Reliable measures of syntactic and lexical complexity: The case of Iris Murdoch. In *Proceedings of the Corpus Linguistics 2017 Conference*, Birmingham, UK.

Literatur

Jarvis, S. (2013). Defining and measuring lexical diversity. S. Jarvis, M. Daller (eds.): *Vocabulary knowledge*. Amsterdam, 13-44

Kornai, András (1999). Zipf's law outside the middle range. In *Proceedings of the Sixth Meeting on Mathematics of Language*, pages 347–356, University of Central Florida.

Kubat, M., J. Milička (2013). Vocabulary Richness. Measures in Genres. *Journal of Quantitative Linguistics* 20, 4, 339-349.

McCarthy, P., S. Jarvis (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42 (2), 381-392.

Tanaka-Ishii, K., S. Aihara. (2015). Computational Constancy Measures of Text. Yule's K and Rényi's Entropy. *Computational Linguistics* 41,3, 481- 502.

Tweedie, F., H. Baayen (1998). How variable may a constant be? *Computers and the Humanities* 32, 323-352.

Yule, G. Udny (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge.